面向数据流的多任务多核在线学习算法 *

裴 乐，刘 群

(重庆邮电大学 计算智能重庆市重点实验室，重庆 400065)

摘 要：多任务多核学习已逐渐成为在线学习算法研究的热点。对于数据流的处理，现有的在线学习算法在准确性上有一定的欠缺，因此提出一种新的多任务多核在线学习模型用于提高数据流预测的准确性。在保持多任务多核学习的基础上，将其扩展到在线学习中，从而得到一个新的在线学习算法；同时为输入数据保持一定大小的数据窗口，用较小空间换取数据的完整性。实验部分对核函数的选取以及训练样本集的大小进行了较为详细的分析，通过对 UCI 数据和实际的机场客流量数据进行分析，很好地保障了流数据处理的准确性及实时性，有一定的实际应用价值。

关键词：多任务多核学习；在线学习；流数据；支持向量机

中图分类号：TP181 doi: 10.3969/j.issn.1001-3695.2017.09.0921

Online learning algorithm based on multi-task and multi-kernel for stream data

Pei Le, Liu Qun

(*Chongqing Key Laboratory of Computational Intelligence*, *Chongqing University of Posts & Telecommunications*, *Chongqing 400065*, *China*)

Abstract: Multi-task and multi-kernel learning has gradually become the research focus of online learning algorithms. For the prediction of data stream, some online learning algorithms have some shortcomings in accuracy. Therefore, this paper proposes a new multi-task and multi-kernel online learning model to improve the accuracy of data stream prediction. Based on the multi-task multiple-kernel learning, we extends the model to online learning, so as to get a new online learning algorithm, while maintaining a certain size of the input data window for the integrity of the data with less space. In the experimental part, the selection of kernel function and the size of training sample set are analyzed in detail. Through the analysis of UCI data and actual airport passenger flow data, the algorithm proposed in this paper can ensure the accuracy and real-time of stream data processing, and has certain applicable value.

Key Words: multi-task and multi-kernel learning; online learning; streaming data; SVM

0 引言

对于许多大数据应用领域，如金融时间序列预测、自然语言处理、网络流量控制等，这些领域中的数据都是实时产生、动态增加的，只要数据源一直处于活动状态，可以认为数据会无限制的增加下去。由于数据比较多，基本上无法全部存储，因此需要在线学习[1]模型进行实时处理。在围绕流数据处理这一需求展开的研究中，在线机器学习算法采用数据流直接处理的模式，每次迭代处理一个随机流数据，学习变量的迭代更新只经过简单的计算，从而在实时性和准确率之间取得一个平衡，是解决该问题很有前途的方案[2]。

经典的在线学习算法有 passive aggressive(PA)方法[3]、感知器算法[4]和基于置信度 confidence-weighted(CW) [5]方法，然而不管是 PA 的超平面模型还是由之衍生的基于置信度的超平面分布模型，都假设数据几乎线性可分，可这个假设不总是成立的，因为数据经常是非线性分布在原始输入空间中，且由于数据采集主观偏差等因素使得数据很难是完美的、符合预设的线性分布，因此进一步研究非线性模型是必不可少的。为了解决在线学习中数据非线性可分的问题，在线多任务学习模型是目前主要研究的一个方向。

在线多任务学习模型旨在同时使用共享信息学习多个相关任务，这样每个任务都可以从学习所有任务中获益。例如文献[6,7]采用混合范数的正则化来为每一个任务学习其权重，这种方式既考虑了任务间的相关性，又在权重的计算方式中实现了任务自身和不同任务间的稀疏性。但其在训练样本较少的情况下很难快速地收敛并训练出相应的模型。为了解决收敛性的问题，Yang [8]等人提出了另外一个简单的信息共享策略，就是让所有的任务共享一个相同的内核函数，这种将特征数据通过核

基金项目：国家重点研发计划资助项目（2016QY01W0200）

作者简介：裴乐（1991-），女，内蒙古巴彦淖尔人，硕士研究生，主要研究方向为多任务在线学习（13618392075@163.com）；刘群（1969-），女，教授，博士，主要研究方向为智能信息处理、在线学习等.

函数的方法映射到另一个特征空间中的方法，可以避免对数据本身进行复杂的运算，同时采用核组合的方法可以更好地得到数据的特征。但是该方法存在的最大问题就是需要进行核函数的选择，不同的核函数选择方式对于结果的影响是十分巨大的。

离线学习方法中的多核学习模型是通过线性或非线性组合几个预先选定的候选核函数来寻求一个适当的核函数，这样可以避免验证不同内核的选择。基于以上分析，本文研究数据流环境下的多任务多核在线学习算法。将多核方法应用于在线多任务学习框架中，通过在线学习以迭代的方式确定一个最适合于当前数据产生规律的组合核函数，避免直接通过数据样本分析对核学习器的更新，使得计算方式更为简单，同时发挥了核方法收敛率高的优点和多任务联合学习的优点，使得数据流的处理更为方便。本文的贡献主要有 2 点：一是将成熟的多任务多核学习框架扩展到在线学习中，形成一个新的算法，该算法可以更好地处理流数据；二是为新产生的数据保持一个数据窗口，在无法及时处理时先保存起来，以保证数据的完整性。

1　多任务多核学习算法

假设有 T 个任务，它们的数据来自于同一个空间 $X\times Y, X\in R^{t}, Y\in R$，这 T 个任务各自拥有不同的数据点 $(x_{i,1},y_{i,1})_{n1},\cdots,(x_{i,T},y_{i,T})_{nT}$，其中 $x_{i,\cdot}\in X, y_{i,\cdot}\in Y$，$n_{t=1}^{T}$ 中是第 t 个任务的数据总量。对于第 t 个任务，它的决策函数为

$$f_t(x)=w_t^T\cdot\Phi(x)+b_t,\forall t\in\{1,\ldots,T\} \tag{1}$$

其中：b_t 是偏置项，$\Phi(x)$ 是将映射后的特征向量，系数 $w_t\propto(w_{t,1},w_{t,2},\cdots,w_{t,M},m=1,2,\cdots,M)$ 是第 t 个任务所对应的所有核系数 $w_{t,m}$ 的集合，并且 w_t 是中每个任务所对应核函数的系数，其中 H_θ 是定义在再生核希尔伯特空间(reproducing kernel hilbert space, RKHS) 中的核函数，$H_\theta=\theta_m\cdot k_{t,m}$，其中 $k_{t,m},m=1,\cdots,M$ 是预先设定的核函数。

多任务多核的目标就是通过最小化经验风险以及权重的正则化项来为每一个任务在有约束的条件下学习一个决策函数，该约束是所有的任务需共享一个共同的稀疏核表示。因此，需要建立一个学习算法能够为每一个任务建立一个函数。为了实现这个目标，本文将问题转换为一个正则化优化问题：

$$\min_{w_1,\ldots w_T}\sum_{t,i}C\cdot L\left(f_t\left(x_{i,t}\right),y_{i,t}\right)+\Omega\left(w_1,\ldots w_T\right) \tag{2}$$

其中：$L(f_t(x_{i,t}),y_{i,t})$ 是损失函数，用于描述数据与模型的契合程度，也就是训练集上的误差；C 是控制模型复杂度与损失惩罚比重的参数；Ω 用于表示不同任务间的关系，是一个包含所有决策系数 w_t 的正则化项，不难看出，如果没有 Ω 约束，则式(2)为 T 个独立学习问题。

2　多任务多核在线学习模型

2.1　算法描述

多任务多核算法本质上来说是一个多目标优化问题，借鉴多目标优化的求解方法，本文通过获得算法的 Pareto 最优解，用标量化的方法来实现该多任务多核问题，也就是通过优化目标函数的不同最优组合，来找到对应问题的最优解。因此问题(2)可看做是一个在多目标优化中的只有一个特定解的 Pareto Front[9]问题，使用参数 λ_t 来标量化问题(2)，即优化问题可重写为

$$\min_{w_1,\ldots w_T}\sum_{t,i}\lambda_t\left(C\cdot L\left(f_t\left(x_{i,t}\right),y_{i,t}\right)+\Omega\left(w_1,\ldots w_T\right)\right) \tag{3}$$

因此，可以通过实现最优化问题(3)来解决问题(2)。本文会有以下一些约定：λ_t 的取值范围为 $\lambda_t\in(1,\ r_\lambda)$，其中 λ_t 是预先给定的。

正则化项 Ω 的基本模型[10]是 $\Omega=\frac{1}{2}\sum\|a\|^2$，这里的 a 表示模型权重系数。本文使用的信息共享策略是学习每个任务的固有特征并且学习所有任务的共同特征，因此模型权重除了每个任务固有的权重 $w_{t,m}$ 外，还有一个包含所有任务共同特征的核系数 θ，所以在讨论权重的正则化项 Ω 时需要同时考虑这 2 个权重系数。Ω 存在的意义就是让各个任务间的差别尽可能的大，使得模型可以更好地进行预测，因此就需要在公共特征的基础上再考虑每个任务的固有特征，根据文献[9]的理论，Ω 表示如下：

$$\Omega\left(w_1,\ldots,w_T\right)=\frac{1}{2}\frac{\sum_m\left\|w_{t,m}\right\|^2}{\theta_m} \tag{4}$$

基于数据流的多任务多核在线学习算法，与批量算法最大的不同之处在于以下两点：a)数据不断产生，无法将所有的数据全部保存再进行处理；b)训练模型需要根据新到来的数据不断更新，以达到更好的预测结果。基于这两点，本文采取以下两种方法依次进行解决。

2.2　引入输入数据窗口概念

在已有的流数据处理算法中，大部分算法都是对新产生的数据逐个进行处理，这就要求算法的实时处理能力很强，如果算法的数据流处理能力不能满足实时性要求，那么部分数据会因为不能及时处理而丢失。为了保证数据的完整性，本文使用一定大小的空间来换取数据的完整性。具体而言，则是算法保持一个固定大小的输入数据窗口，每个窗口可保存一个数据样本，在新数据不断产生时，将无法及时处理的新样本数据保存到该窗口中，以免因为未及时处理而丢失数据，并且一定程度上降低了对算法实时性能的要求。对于本文所使用的基本模型即 SVM 算法来说，输入样本数是一个或多个不会影响算法的整体运行速度，因此保持一个输入数据窗口是很有必要的，本文通过实验证明该窗口数为 3 个时效果可以达到最优。

2.3　模型更新策略

对于模型更新问题，基于流数据无限产生的特点，每次预测错误都会增加一个新的支持向量，这样看来支持向量的数目没有上界的，直接计算整个数据集的核矩阵 K 在计算资源上并不现实。因此本文通过给定一个支持向量的最大上限来解决该

问题，这是在牺牲一定的准确率基础上实现的，后面的实验中详细讨论了该最大数目设置对于预测结果的影响。对于SVM来说，算法需要保持一个固定大小的训练样本集来进行核矩阵的计算，引入在线学习技术后，该样本集中的数据则是需要保存的支持向量的数据，因此必须有一个策略对训练样本集中的数据样本进行更新。

本文借鉴操作系统中的内存页面的调度算法[11]，采用先进先出(first in first out, FIFO)策略对数据样本进行更新，对于数据流，其数据生成规律有可能随时间变化而变化，因此替换存在时间最长样本的FIFO策略是合理的。根据FIFO策略，对训练样本集的更新如下：每次把新加入的样本放在最下行和最右列，然后去掉第1行和第1列即可完成训练样本集的更新[12]。这种限制工作集大小的更新策略有一定的局限性，但在有限的计算和存储资源下是折中的策略。

另外在更新训练模型之前，本文预先对公共核系数进行更新，对每个预测任务进行惩罚，减少模型训练的迭代系数，进而降低训练时间。在第 j+1 次迭代时，θ 的更新公式为 $\theta_m^{j+1}=\theta_m^j\cdot\beta$，β其中为一个0到1之间的随机数。

算法1较为详细的描述了该模型的整体过程。

算法1　多任务多核在线学习算法

输入　训练样本集：$D_t^j=\{x_1,x_2,...,x_N\}$;核函数集合：$K_m=\{k_1,k_2,...,k_M\}$;依次给每个任务输入一个样本 x_0 或一组样本 D_0。

输出　预测结果和更新后的模型。

根据新来的样本为该任务计算相应的核矩阵；

获得已有的该任务训练模型 $Model_t^j$；

利用模型 $Model_t^j$ 预测结果；

if 预测值==真实值

　接收下一个数据

else

　用 x_0 更新训练样本集的最后一行和一列

$$D_t^{j+1}=\{x_2,x_3,...,x_N,x_0\};$$

更新核组合系数 $\theta_m^{j+1}=\theta_m^j\cdot\beta$；

根据SVM算法，用 D_t^{j+1} 和 θ_m^{j+1} 为该任务训练新模型 $Model_t^{j+1}$。

算法1对本文提出的在线学习算法进行了整体的描述，为了处理不断到来的流数据，把多任务多核学习框架和在线技术进行结合生成一个新的算法，在每次处理一个或多个新到来的数据的基础上，同时将数据样本集不断的进行更新，可以更好的得到各个任务的模型，从而更好的进行结果预测。

2.4　时空代价分析

大数据流的数据规模大，到达速率非常快，要求数据流挖掘算法在有限的内存空间中实时处理，这就要求面向大数据流分析算法的时间、空间复杂度低。由于本文在输入数据时保持了一个输入数据窗口，因此对于算法的实时性要求相应的降低一些。下面对算法的时空复杂度做了一个简单的分析。

假设有Q个任务，每个任务有k个d维的训练样本，其总共有N个训练样本 $(N\succ k)$，有m个核函数，对于本文所提出的在线学习算法MTMKOL来说，每次迭代的时间复杂度为 $O(Q\cdot m)$，当样本顺次出现时，时间复杂度为 $O(Q\cdot m\cdot N)$。

对于空间复杂度来说，算法需要一直保存一个训练样本集，其大小为 $O(Q\cdot d\cdot k)$，同时要保持一个大小为3的输入数据窗口，其所需的空间大小为 $O(Q\cdot d\cdot 3)$，所以本算法的空间复杂度为 $O(Q\cdot m\cdot k+Q\cdot d\cdot 3)=O(Qmk)$。

而算法ADA-MTL[7]的时间复杂度为 $O(QdN)$，算法BMKOL[13]的时间复杂度为 $O(QkN)$，而批处理算法MTMKL[14]的时间复杂度为 $O(QmNN)$。

根据以上分析可知，本文提出的算法与其他两个在线学习算法的时间复杂度都是 $O(N)$，满足在线学习算法时间复杂度的要求。

2.5　核函数的选择

对于多核学习问题来说，如何选择核的个数及种类在目前来说没有一个统一的理论选择标准，因此“核函数选择”成为支持向量机的最大变数。本文通过对常用的核函数进行一个简单的分析，并通过实验来选择本文中所使用到的核函数。

常用的核函数有多项式核、线性核和高斯核。本文中选取1个多项式核、1个线性核和多个高斯核，其中高斯核的数量的通过实验来进行确定。本实验使用同一样本集 robot 来验证，其中输入数据窗口数为3，训练样本集的大小为总样本数的10%。表1展示了在多项式核、线性核一定的情况下，不同高斯核个数所对应的分类准确率以及所需要的运行时间，其中分类准确率是预测正确样本数占总样本数的比重，运行时间是整个数据集全部实验完成的时间。表中所有的值都是运行10次后所取的平均值。

表1　不同高斯核个数所对应分类及运行时间情况

多项式	线性核	高斯核	准确率	运行时间
1	1	1	0.9778	3.9777
1	1	3	0.9694	4.3414
1	1	5	0.9778	4.0698
1	1	7	0.9803	3.4041
1	1	9	0.9736	4.1452

从表1可以看出，高斯核个数的不同，所对应的分类预测准确率和运行时间都有所不同，其中效果最好的是当有5个高斯核时，其运行时间最少并且其分类准确率也达到了0.98，结果都是最优的，因此本文中的最优核函数组合的选取为1个多项式核、1个线性核和5个高斯核的基本核函数组合。

3　实验及结果分析

在本节中，本文分析比较的方法包括本文提出的多任务多核在线学习算法MTMKOL、多任务加速在线学习算法[7]ADA-MTL、基于预算量的多核在线学习算法[13]BMKOL和多任务多

核学习算法[14] MTMKL，其中表 2 给出对比实验算法的相关信息。为了比较本文提出算法的可伸缩性能，本文使用了 UCI 数据集[15]中的 robot 和 letter 数据集来进行实验，通过讨论算法中不同训练集大小、不同输入数据窗口大小的使用，得到本算法的最优训练集合大小以及最优输入数据窗口大小，并且通过对比讨论不同算法，验证了本文算法较好地伸缩性能。为了进一步验证本文算法在实际数据流场景下的使用性能，本文使用阿里天池广州白云机场客流量预测[17]的数据，验证了本算法和其他三个算法在选定数据集上的有效性，同时得到不同任务数量所对应的算法效率情况，进而对算法的处理能力有整体的认识。本文所有实验均在 MATLAB R2010b 上运行。

表 2　实验对比算法的基本信息

缩写	参考文献	算法描述
ADA-MTL	7	基于正则化对偶平均方法的多任务在线学习算法，结合使用加速技术提高收敛速度
BMKOL	13	基于预算量的多核在线学习算法，更新最小二乘支持向量机来进行预测
MTMKL	14	多任务多核学习算法

3.1　UCI 数据实验分析

为了评估本文在线学习算法的分类性能，本文从 UCI 数据集中选取了 2 个多任务数据集来进行实验，其中这两个数据集为 robot 和 letter，robot 有 6 个任务，每个任务有 500 个样本，并且每个任务有 4 个属性；letter 是手写体单词（8 个任务），每个任务有 500 个样本，并且每个任务有 16 个属性。本小节使用分类准确率和运行时间 2 个指标来评估算法的性能。其中分类准确率是预测正确样本数占总样本数的比重，运行时间是整个数据集全部实验完成的时间。

3.1.1 算法训练样本数选择

本文的算法输入要求保持一个大小为 $Q \cdot N \cdot d$ 的存储空间来存放训练样本集，其中 N 的取值直接决定算法所需要占用的内存空间。为了得到最优 N 的取值，本文选择多任务数据集 robot 进行实验，其中输入数据窗口为 3 个。图 1 比较了不同训练样本集大小所对应的分类准确率以及所需要的运行时间，图中所有的值都是分别进行 10 次随机实验的平均值。

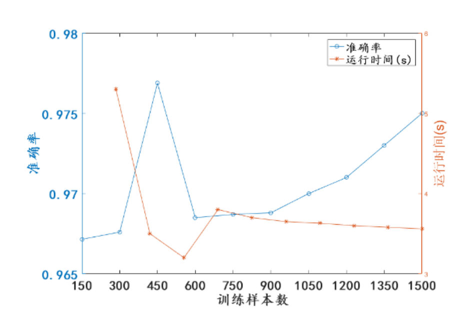


图 1　不同训练样本数所对应的准确率和运行时间

如图 1 所示，随着训练样本个数的增加，分类预测准确率整体上在不断的提高，同时实验所需要的运行时间整体上在降低。其中效果最好的是当样本数为 75 时，它对应的分类准确率和运行时间都是最优的，同时样本数量在 50-75 之间时，准确率和运行时间都比较好，因此本文所有的实验设置中训练样本集的大小都在 50-75 之间取值。上图中分类准确率高达 0.99，其原因是支持向量机本身对分类问题处理效果非常好，再加上不断对模型进行更新与修正，使得错误率大幅度下降。

3.1.2 算法输入窗口数选择

本文的算法输入要求保持一个大小为 $Q \cdot N \cdot d$ 的存储空间作为输入数据窗口，其中 n 的取值影响算法所需要占用的内存空间。为了得到最优 n 的取值，本文选择多任务数据集 robot 进行实验，其中训练样本个数为 75。图 2 比较了不同窗口大小所对应的预测结果，图中所有的值都是分别进行 10 次随机实验的平均值。

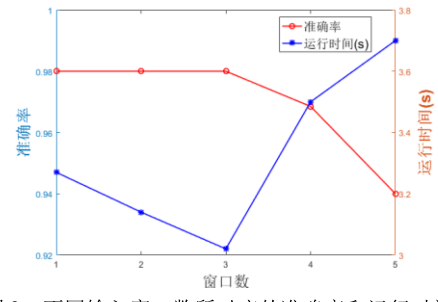


图 2　不同输入窗口数所对应的准确率和运行时间

如图 2 所示，随着输入窗口个数的增加，分类预测准确率先保持不变然后又突然下降，同时实验所需要的运行时间先减少然后又突然增加。其中效果最好的是窗口数为 3 时，它对应的分类准确率和运行时间都是最优的，因此本文所有的实验设置中输入窗口都为 3。上图中窗口数为 3 是一个拐点，其原因是窗口个数过大时，容易包含预测错误的样本，而程序需要从当前窗口的所有样本中找到出错的样本再进行模型的更新，这个过程会产生额外的时间开销，同时准确率也会下降。

3.1.3 算法可伸缩性分析

为了评估本文流数据在线算法的可伸缩性，本文选择 2 个多任务数据集 robot 和 letter 进行实验，其中输入窗口个数为 3。实验设置如下：训练样本分别占总样本不同比例，即{5%，10%，15%，20%，25%，30%，35%，40%，45%，50%}，剩余的样本作为评估使用。图 3 分别比较了各种不同设置下本文提出的算法与其他三个对比算法的预测结果。

图 3(a)图是使用 robot 数据集进行实验的，(b)图是使用 letter 数据集进行实验。从图中可以得到以下一些结论：

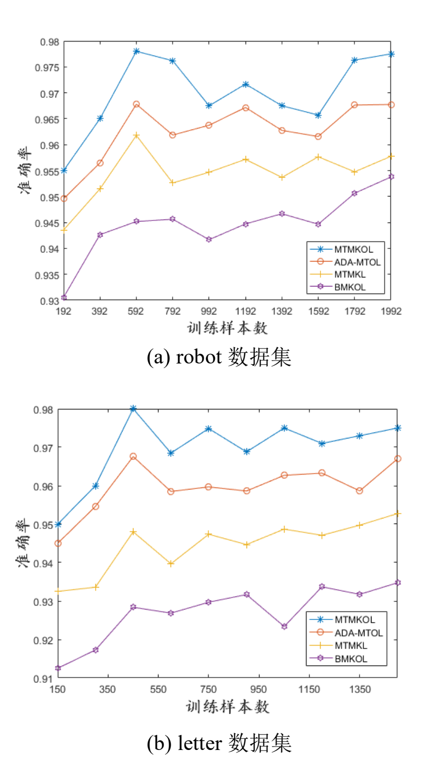
a）本文提出的算法 MTMKOL，可以明显的看出其分类的准确率一直大于其他三个算法，并且其分类的准确率也非常高。这是因为一方面本算法选用的支持向量机这个模型就是一个非常善于分类的模型；另一方面本算法维持了一个输入集，对实时性的要求相对减少，因此算法主要是从准确率这一方面进行考虑的。并且本文所提出的算法在不同训练样本数量的设置下，所得到的准确率值差值很小，进一步说明了该算法对训练集中

的样本数量要求很低，具有很好地可伸缩性，可以满足大规模数据流的需求；

b）算法 BLLSVM 的分类准确率明显低于其他三个算法的准确率，其原因可能是该算法对支持向量的个数有一定的限制，其次它对算法实时性的要求是以牺牲一部分的准确率作为代价的，因此它的准确率稍低；

c）算法 ADA-MTL 的分类情况不是很稳定，其原因是算法不需要保持一定数量的历史数据，直接计算权重对结果会有一定的影响；

d）算法 conic MTL 是一个批量处理的算法，根据所给训练样本学习出的模型用于对新数据的预测，其结果会随着训练样本数的增加而更加准确。



(a) robot 数据集

(b) letter 数据集

图 3　各种设置下算法的准确率比较

3.2　阿里天池机场客流量流数据实验分析

为了评估本文在线学习算法在实际应用中的回归性能，将使用机场客流量的时空分布预测[16]数据来进行实验，根据其在机场客流量预测中的预测误差 RMSE(root mean square error)与数据吞吐量来衡量模型与算法性能。从而验证 MTMKOL 在机场客流量预测中的特性与优势。

数据集是来自天池竞赛机场客流量的时空分布预测的初赛数据集，其提供海量机场 WIFI 数据及安检登机值机数据，以对白云机场航站楼客流量分析与预测。初始数据集中的时间数据都是精确到秒，将其进一步简化，以 10 min 作为一个时间片，将数据重新进行汇总保存。将每一个 WIFI 点的客流量预测看做是一个任务，则共有 749 个任务。对于每一个任务来说，机场每天的排班表基本稳定，用户在机场内的行走模式也基本稳定，并且时间序列具有一定程度的连续性，某一时间点的情况会一定程度延续此前几小时的情况和前 2 天的情况，因此使用不同时间的客流人数作为任务的数据特征，具体信息如表 3 所示。

表 3　数据特征信息表

编号	特征名称
1	前 10 分钟人数
2	前 20 分钟人数
3	前 1 天该时间点人数
4	前 2 天该时间点人数

3.2.1 算法回归性能分析

为了评估本文流数据在线算法在回归问题中的性能，本文选取数据集中的前 100 个任务点来进行实验。其中训练集选取了 12 个小时的数据，即 72 个数据样本，测试集为剩余的 1510 个数据样本。图 4 比较了不同算法在该数据集上回归分析的预测误差 RMSE，其中预测误差 RMSE 是真实值与预测值误差的平均平方根。

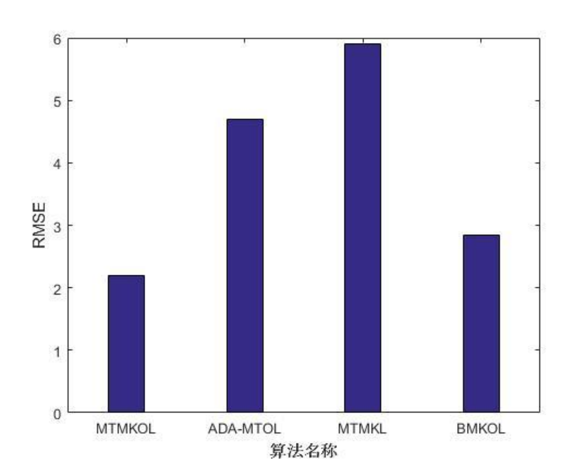


图 4　选定数据集上回归分析的 RMSE 值

如图 4 所示，不同算法所对应的 RMSE 是各不相同的。其中本文提出的算法 MTMKOL 的 RMSE 是最小的，主要原因是使用了 SVM 算法，它可以使用很少的数据训练出较好的模型，使得预测误差值变得很小；而批处理算法 Conic MTL 的误差值是最大的，其原因是算法只训练了一次模型就进行测试，但是该模型并没有很好的匹配数据特征，从而导致了较大的误差值；其他两个算法的误差值也比较大一点，原因则是给定训练样本集较小，刚开始模型并不能很好的匹配数据，经过不断的更新模型才使得误差值逐渐减小。

3.2.2 算法不同任务数回归分析

为了评估本文流数据在线算法在不同任务数量中的回归性能，选取数据集中所有的 749 个任务点进行实验。实验设置如下：任务数量为{2,100,200,300,400,500,600,749}，分别验证不同任务量的情况下算法预测误差值以及数据吞吐量，其中训练集选取了 12 个小时的数据，即 72 个数据样本，测试集为剩余的 1510 个数据样本。图 5 比较了不同任务个数在该数据集上回归分析的预测误差 RMSE 和数据吞吐量，其中数据吞吐量是每秒钟所能处理的数据量。

从图 5 中可以观察到下列现象：

多任务学习方式比各个任务单独学习方式的回归预测性能要好许多，并且随着任务数的增加，相对应的 RMSE 和吞吐量有所优化；

任务个数越多，其相应的预测误差值越小，预测值更接近真实值；

随着任务个数的增多，其每秒钟可以处理的数据不断增加，算法的吞吐量最高可达到每秒 320 个数据样本，完全满足了算法实时性的要求。

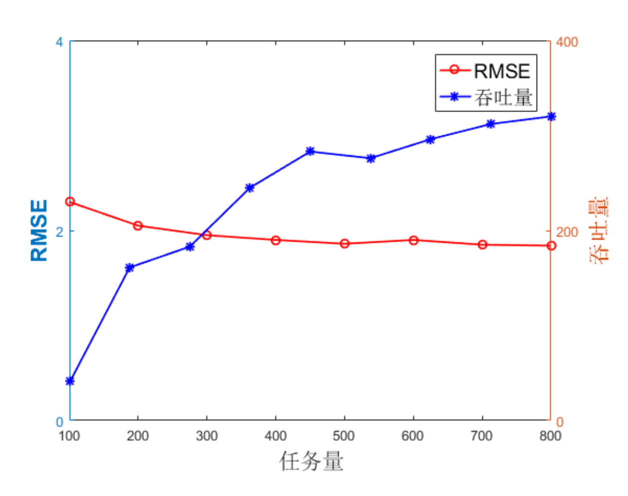


图 5　算法 MTMKOL 预测性能与任务个数的关系

4　结束语

本文提出一种新颖的面向数据流的多任务多核在线算法，首先通过实验对最优核函数组合进行选取，由于现如今对核函数的选择没有一个统一的理论标准，因此本文根据已有的经验进行实验选择；其次通过实验对输入样本集大小进行选取，使得预测准确率与所耗费时间都能得到较好的结果，同时与现有的在线算法和批处理算法进行比较，本算法在各种实验中均取得更好的效果；然后，已有的在线学习算法都是直接对新到来的数据进行处理，算法需要很高的实时处理能力，而本文提出的算法则是为新到来的数据保持一个数据输入窗口，可以很好地避免了数据的丢失以及降低对算法的实时性要求，实验结果表明这样的设置是非常有意义的；最后对真实的机场客流量数据进行分析预测，结果得到预期效果。

本算法比较关注对流数据的处理，在核系数的更新算法中处理比较简单，本文将在接下来的工作中仔细研究该更新算法，使得模型的更新更为简单快速。

参考文献：

[1] 李志杰，李元香，王峰，等. 面向大数据分析的在线学习算法综述 [J]. 计算机研究与发展, 2015, 52 (8): 1707-1721.

[2] 潘志松，唐斯琪，邱俊洋，等. 在线学习算法综述 [J]. 数据采集与处理, 2016, 31 (6): 1067-1082.

[3] Wang Z, Vucetic S. Online passive-aggressive algorithms on a Budget [J]. Journal of Machine Learning Research, 2010, 9 (9): 908-915.

[4] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. [J]. Psychological Review, 1958, 65 (6): 386.

[5] Crammer K, Dredze M, Pereira F. Confidence-weighted linear classification for text categorization [J]. Journal of Machine Learning Research, 2012, 13 (1): 1891-1926.

[6] Rakotomamonjya, Flamary R, Gasso G, et al. lp-lq penalty for sparse linear and sparse multiple kernel multitask learning [J]. IEEE Trans on Neural Networks, 2011, 22 (8): 1307-20.

[7] 李志杰，李元香，王峰，等. 面向大数据流的多任务加速在线学习算法 [J]. 计算机研究与发展, 2015, 52 (11): 2545-2554.

[8] Yang H, Lyu M R, King I. Efficient online learning for multitask feature selection [J]. ACM Trans on Knowledge Discovery from Data, 2013, 7 (2): 1693-1696.

[9] Li C, Georgiopoulos M, Anagnostopoulos G C. Pareto-path multitask multiple kernel learning [J]. IEEE Trans on Neural Networks & Learning Systems, 2015, 26 (1): 51.

[10] 周志华，王珏. 机器学习及其应用 [M]. 北京：清华大学出版社, 2007.

[11] 邹恒明. 计算机的心智：操作系统之哲学原理 [M]. 北京：机械工业出版社, 2012: 100-102.

[12] 张钢，谢晓珊，黄英，等. 面向大数据流的半监督在线多核学习算法 [J]. 智能系统学报, 2014, 9 (3): 355-363.

[13] Jian L, Shen S, Li J, et al. Budget online learning algorithm for least squares SVM [J]. IEEE Trans on Neural Networks & Learning Systems, 2016, 28 (9): 2076-2087.

[14] Li C, Georgiopoulos M, Anagnostopoulos G C. Conic multi-task classification [C]// Machine Learning and Knowledge Discovery in Databases. 2014: 193-208.

[15] UCI 数据集 [EB/OL] http://archive. ics. uci. edu/ml/DOI.

[16] 机场客流量时空分布预测 [EB/OL]. https://tianchi. aliyun. com/competition/introduction.htm?spm=5176.100066.333.4.6YizCQ&raceId=231588DOI.